Heliyon

ELSEVIER

Received:
18 July 2016
Revised:
4 September 2016
Accepted:
30 September 2016

Heliyon 2 (2016) e00175

CrossMark

Melody discrimination and protein fold classification

Robert P. Bywater [a], **Jonathan N. Middleton** [b,c,*]

[a] *Francis Crick Institute, London NW1 1AT, UK*

[b] *Department of Music, Eastern Washington University, Cheney, WA 99004, USA*

[c] *School of Information Sciences, University of Tampere, 33041, Finland*

* Corresponding author.

E-mail address: jonathan.middleton@uta.fi (J.N. Middleton).

Abstract

One of the greatest challenges in theoretical biophysics and bioinformatics is the identification of protein folds from sequence data. This can be regarded as a pattern recognition problem. In this paper we report the use of a melody generation software where the inputs are derived from calculations of evolutionary information, secondary structure, flexibility, hydropathy and solvent accessibility from multiple sequence alignment data. The melodies so generated are derived from the sequence, and by inference, of the fold, in ways that give each fold a sound representation that may facilitate analysis, recognition, or comparison with other sequences.

Keyword: Bioinformatics

1. Introduction

Globular proteins are linear copolymers of amino acid residues that have stretches of more or less regular geometry which are packed together in ways that at first sight appear anything but systematic. The locally regular regions, called secondary structures, are either helical (α- or 3_{10} helix type) or β-strands, which can arrange themselves in either a mutually parallel or antiparallel fashion to form so-called β-sheets. Each amino acid residue type has a different propensity to favour one or other of these structural arrangements [1, 2, 3, 4]. The fold that ensues is likewise dependent on another property of the amino acids, the hydropathy, or degree to

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

which the side chain of the residue seeks or avoids contact with the solvent water. This in turn affects the degree to which a given residue will be buried in the interior or lie on the surface. These are three of the major attributes of polypeptide sequences that contrive to determine the final 3-dimensional (3D) structure, and all three can be determined by present-day bioinformatics methodology (see Methods). One more complication that needs to be dealt with is the fact that the polypeptides themselves may fold into more than one "independently folding unit" along the chain. These units are referred to as domains. For small proteins there is typically only a single domain, but larger proteins may have two or more. These can be similar in character (and as such may be the result of a gene duplication event) or different, depending on the functional needs of the protein. The assembly process, which is well understood at the level of molecular biology, enlists a catalytic process referred to as splicing. This involves selecting disjoint sections from the original DNA sequence (genome) for translation into the "language" of proteins whereafter the splicing process takes place in a defined order. Protein folds, or more correctly, domain folds, can be classified in one of several ways [5, 6, 7]. With this as a background to the pattern recognition problem, we now describe the approach we are taking towards its solution.

Here, we explore the use of melodic sonification to discern patterns of data that are based on certain features associated with protein structures: evolutionary information, chemical propensities and physical attributes. We anticipate that the perception of protein data via sonic representation can assist researchers in the process of pattern recognition and structural understanding. Melodic patterns of sound can become tools to assist protein chemists in assigning fold types to a given protein with a known sequence, but unknown 3D structure.

The potential benefits for using auditory display of scientific data as a means to derive analytical meaning have been formally demonstrated since 1994 [8, 9, 10]. These authors argued that the human auditory system has the sophistication and the ability to interpret sounds "using multiple layers of understanding" [10]. In reference to the choice of analytical listening over visual review of data, suggestions have been made [11] to the effect that "patterns may emerge which are otherwise undetectable". Although the analytical and interpretative processes (through listening) are not completely understood, the perceptual aptitudes have recognized merits: data-to-sound parameter mappings and applications now serve many disciplines such as "chaos theory, bio-medicine, interfaces for visually disabled people, data mining, and seismology" [10]. More specifically to genome science, we anticipate that sonifications could highlight patterns to find mutants and genetic disease markers for which a rich collection of sequence data is available.

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

While most data are observed, sonifications that are designed specifically for coherent data-to-sound representation can complement more traditional visual displays of data and offer a useful secondary perceptualization to analyze data sets. In this paper we seek to create sonifications of protein structures that can uniquely enhance our ability to recognize patterns related to protein folds. We have chosen pitch as the most basic and primary sonification element to represent data; hence the data sets are converted into sequences of pitches. The conversion results are melodic in nature, demonstrating a simple and straightforward conversion to clearly perceive the sonified data patterns. Other sound elements such as harmony and counterpoint with timbral variety did not yield simple listening environments for sequential data pattern recognition. In some cases rhythmic mapping was applied as a secondary sonification parameter to enhance the melodies.

The present study contributes to a relatively small collection of previously explored sonifications of protein structures. In many cases, previous studies sought to educate and inspire readers with interdisciplinary connections between music and science, and making science more widely accessible. Efforts, such as the *Life Music* project, were designed for dual purposes of creating new perceptual models for protein analysis and enhancing electro-acoustic music compositions [12, 13, 14, 15]. In each sonification study for proteins, the projects have utilized their own set of mapping systems to convert data-to-music. As an example, the *Life Music* project [12] mapped protein residues to pitches via hydropathy scales, a concept that is similar to one of the mapping systems explained later in this paper (with results that are quite different). In the Takahashi and Miller pilot study [13] amino acids were assigned to discrete pitches, and then to chords and rhythm to make the results more musical.

2. Methods

2.1. Data preparation from protein sequence databases

The sonification study reported at this time is featuring three proteins, each representing three SCOP [5] classes, α-helix type, proteins composed of β-strands, and mixed α/β type.

α-helix: 1ny9.pdb (MW 10.60 kD, α-helix 73.4%, 3_{10}-helix 3.2%, β-strand 0.0%, turn 10.6%, coil 12.8%)

β-sheet: 1r75.pdb (MW 7.85 kD, α-helix 0.0%, 3_{10}-helix 0.0%, β-strand 23.5%, turn 26.5%, coil 50.0%)

α/β type: 4ake.pdb (MW 23.59 kD, α-helix 43.5%, 3_{10}-helix 5.1%, β-strand 15.4%, turn 20.6%, coil 15.4%)

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

In preparation for our overall study, we originally chose proteins identical to representatives of these domain fold classes studied in a recent publication [16] that had as its aim the investigation of domain fold space. The proteins chosen were 1ny9, 1r75, and 1div.pdb (MW 16.31 kD, α-helix 32.2%, 3_{10}-helix 6.0%, β-strand 30.9%, turn 13.4%, coil 17.4%). Since these proteins are relatively small (sequence length, MW), a fourth protein, 4ake, was added to the study. Like 1div, 4ake is a α/β type, but much larger. Further, for the present study, we decided to exclude 1div in lieu of 4ake, although the 1div protein features in an extension of this work, which is already underway. The reason for excluding this protein at this stage has to do with the structural parameters chosen for the first part of this work (see below).

In the case of 1ny9, this is an NMR structure containing 10 alternative solutions to the structure. The best consensus structure, no. 8, was determined using the WHAT IF program [17]. The selected structure herein simply called 1ny9 is, strictly speaking, the structure with PDB identity number 1ny9 having chain identifier A and is the 8th structure in the NMR ensemble. This was chosen on the grounds that it was the closest to the average structure in that ensemble and therefore deemed to be most representative.

The sequence of each of these proteins was then submitted to the PredictProtein server [18]. Predicted structural parameters were extracted from the output files from this server (this was done manually but an automated version is under development). The structural parameters included in our schemes were:

- SeqNo - Sequence number for at the position under consideration
- AA - identity of the amino acid residue at the position
- OHEL - Observed secondary structure
- PHEL - Predicted secondary structure
- OACC - Observed solvent accessibility
- PACC - Predicted solvent accessibility

The above are all obtained directly as outputs from PredictProtein. OHEL and PHEL employ the HST terminology [19] whereby H stands for α-helix, S for β-strand and T for turn. In practice we employed the augmented set defined within the WHAT IF program whereby H, S, and T are defined as before [19] and important secondary structure features not included in HST are added: 3 for 3_{10}-helix and C for coil. The importance of distinguishing C from T is that the former has the character of "random coil" while turns conform to stringent limits on backbone torsion angles. N.B. in much of the bioinformatics literature E ("extended") is used instead of S. For this work, a python script was written to

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

extract these data but there is every likelihood that this will be fully automated in the near future.

While the PredictProtein program always returns a PHEL and PACC value, it does not always give the corresponding OHEL and OACC values. In such cases, the corresponding observed values can be determined using WHAT IF, or other protein modelling software. In our case, we were only interested in OHEL and OACC in training examples. When a protein with unknown crystal structure is studied the data will by definition not be available. It is for this reason that PHEL and PACC are mentioned, since there may be many reasons why researchers might want to study proteins that lack a crystal structure (mutants for example).

Finally, we included structure information not generated by PredictProtein:

AAH – Amino-acid hydropathy – values for each of the 20 amino acid residue types are obtained from a database [20] that provides the currently most reliable set of data.

The hydropathy data are an important feature of the overall corpus of protein structural information, but there is a problem in that values for certain pairs of amino acid types, for example Asp and Glu, and Arg and Lys, are very close to one another, which rather masks the distinctly different structural propensities that these residue types have (see Table 1).

2.2. Sonification of protein sequence and property data

The input data for the three proteins selected for this study are listed in supplementary tables 1–3. The columns contain both experimentally derived data and data derived from prediction methods. The data were converted into a format that can be assimilated by the **musicalgorithms** Web-based software [21].

2.3. Data sonification mapping

In order to discern data patterns most clearly as melodic expressions, the data-to-music mappings focused primarily on pitch with uniform rhythmic durations, and pitch with some rhythmic variety. Future investigations with musical refinements should include secondary characteristics such as timbre, dynamics, and articulations.

The data-to-music mapping of amino acid chains was based on a proportionate expansion algorithm that transforms amino acid values to fit a musical range. The amino acid values from supplementary table 1 were drawn from corresponding whole-residue hydrophobicity measurements (after averaging) within the water to octanol range obtained from the blanco.biomol.uci.edu Web site, and listed in Table 1. The mapping results can be displayed as a musical scale ascending from

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

Table 1. AA hydropathy scale with averaging; adapted from original data [20] and modified for enhanced discrimination.

AA	AA-Hydropathy Averages
W	-2.1
F	-1.7
L	-1.3
I	-1.1
Y	-0.7
M	-0.7
V	-0.5
C	0.0
P	0.1
T	0.3
S	0.5
A	0.5
Q	0.8
N	0.9
G	1.2
H	1.2
R	1.8
E	1.9
D	2.0
K	2.8

aromatic to charged types with 17 discrete pitches assigned to the range of 20 amino acids (see Fig. 1 and supplementary Audio 1). Note, in three cases amino acids were paired up to share a single pitch: M and Y, G and H, and A and S. There is no particular *a priori* rationale for these pairings, but there are some possible

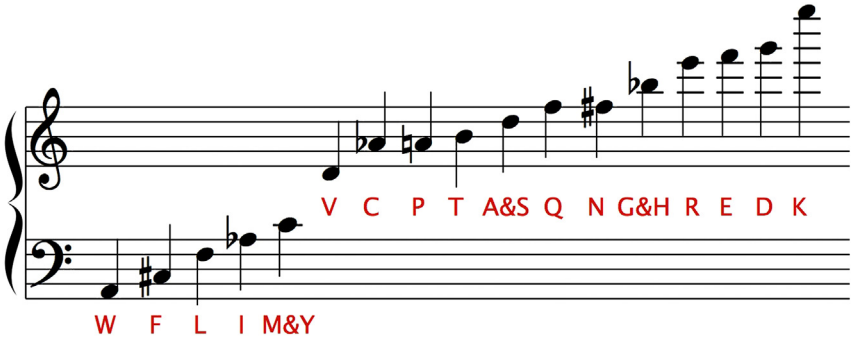

Fig. 1. Musical scale created from AA whole-residue hydrophobicity measurements (after averaging).

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

biophysical explanations for the M/Y and A/S pairs. The latter residue types both represent the "small side-chain" class of residues and they can indeed "replace" each other – examples of entries in multiple sequence alignments that read something like "AAAAASSSASAAAAASSAAAA" are very common. The M/Y case is a bit more subtle. M is similar to Y but also to F and W. They are all members of the "bulky side-chain" group and M can in many instances replace the other three. The case of G and H, these also need some explaining since H is thought of in terms of its highly polar and ~50% (at physiological pH) charged side chain while G has no side chain. On the scale used here, these two lie very close together, but that is also the case with several other hydropathy scales. Most of the experimental techniques used to determine these scales rely in one way or another on partitioning between two phases. We have to accept the fact that H distributes itself across these two phases in much the same way as does G. In the aqueous phase the relative population of the charged species would be determined by the pH, while in the nonpolar phase only the uncharged species would be present.

A proportionate expansion algorithm was used to distribute protein values within a designated pitch range as defined by the user. There is flexibility in setting the destination pitch span to wide or narrow ranges, so that users can shape the musical results according to listener preferences. In any setting, the results will demonstrate some predictability, for example, Trp (W), with an average hydropathy value of −2.09, will always anchor the lowest end of the chosen pitch range; Lys (K), with an average hydropathy value of 2.80, will always map to the highest end of the pitch range; and Ser (S), with an average hydropathy value of 0.46 will map to the middle of the pitch range. In this study, the AA values were mapped to a pitch span of 55 musical notes: piano keys 25–80 (see Fig. 2 and supplementary Audio 2). This setting provided a relatively even distribution of independent pitches for 17 AA values. The same mapping distribution can apply to a range of musical durations. In these cases a set of musical durations from short to long can be associated with numerical data values from low to high; however, no rhythmic variety was added at this time so that the complexities of the 17-pitch scale could be presented in a simplified form. Hence, each pitch in the melodic output was represented by a uniform rhythmic value.

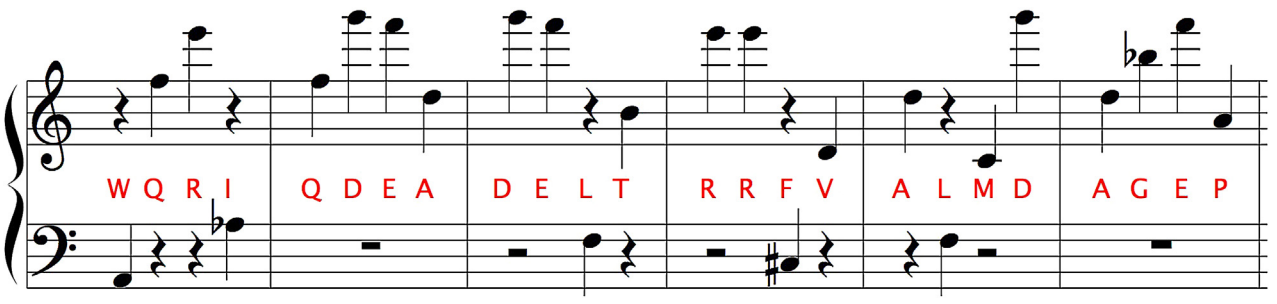

Fig. 2. Introductory excerpt from 1ny9 score representing AA hydrophobicity values.

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

The data-to-music mapping for OHEL values offered the challenge of converting shapes into music. This approach required a preliminary step that transformed OHEL data into "shaped" numeric sequences, which were then sonified to short musical gestures – called motives and sub-phrases:

Coil (C) = 1.0, 0.5, 0.8, 0.5, 1.0, 1.5, 1.2

Turn (T) = 2.3, 1.7, 1.7, 2.0, 2.3

3_{10}-helix (3) = 3.0, 2.8, 3.0, 3.2

α-helix (H) = 2.0, 1.6, 1.8, 2.2, 2.4

β-strand (S) or Extended (E) = 6.0, 5.8

The numeric representation for each shape can be flexibly mapped with the proportionate expansion algorithm to a variety of musical ranges that will expand or compress the shapes on a vertical pitch space (or axis). In music we call this augmenting or diminishing the intervals, which are the distances between pitches. The choice of musical range, whether by expansion or compression, can also alter the general tonality of the results and the size of the rhythmic durations. Hence the choice of musical range should be carefully considered to meet the goals of the analysis and listener preferences to encourage repeated listening.

The OHEL protein data for 1r75 consist of a sequence of shapes labeled C, H, T, S. Each protein residue shape is mirrored by a musical motive of similar contour. In this case the sonification of one data point (or shape) will feature a short collection of pitches rather than a single pitch as seen in the previous mapping for AA hydropathy. For example, the T (or turn) might be represented by the nearly conjunct musical motive: Ab-F-F-F#-Ab. This descending and ascending musical gesture has a U shape, which mirrors the geometry of the T structure (dihedral angles arranged so that the chain reverses). As another example, 3_{10}, a repeating regular structure, is represented musically by repeating pitches as if in oscillation, such as: B-Bb-B-C. In this study each OHEL shape from 1r75 is represented by a motivic pattern of pitches and with varied rhythms that adhere to the same proportionate mapping system (see Fig. 3 and supplementary Audio 3), where large values such as 6.0 generate long rhythmic durations and small values such as 0.5 become short rhythmic durations.

The OACC values correspond to surface and depth relations of the interior of the protein structure. The OACC data points are mapped to the full range of the piano to capture the varying surface-to-depth relationships. In this melodic sonification, pitch is the only point of focus as an effective musical characteristic to track data surface to depth relations. In this example, depth is represented by the deep sounding low pitch content, and surface areas by high pitch content. The uniform rhythmic durations at a steady tempo, or pulse, help the listener focus on essential

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

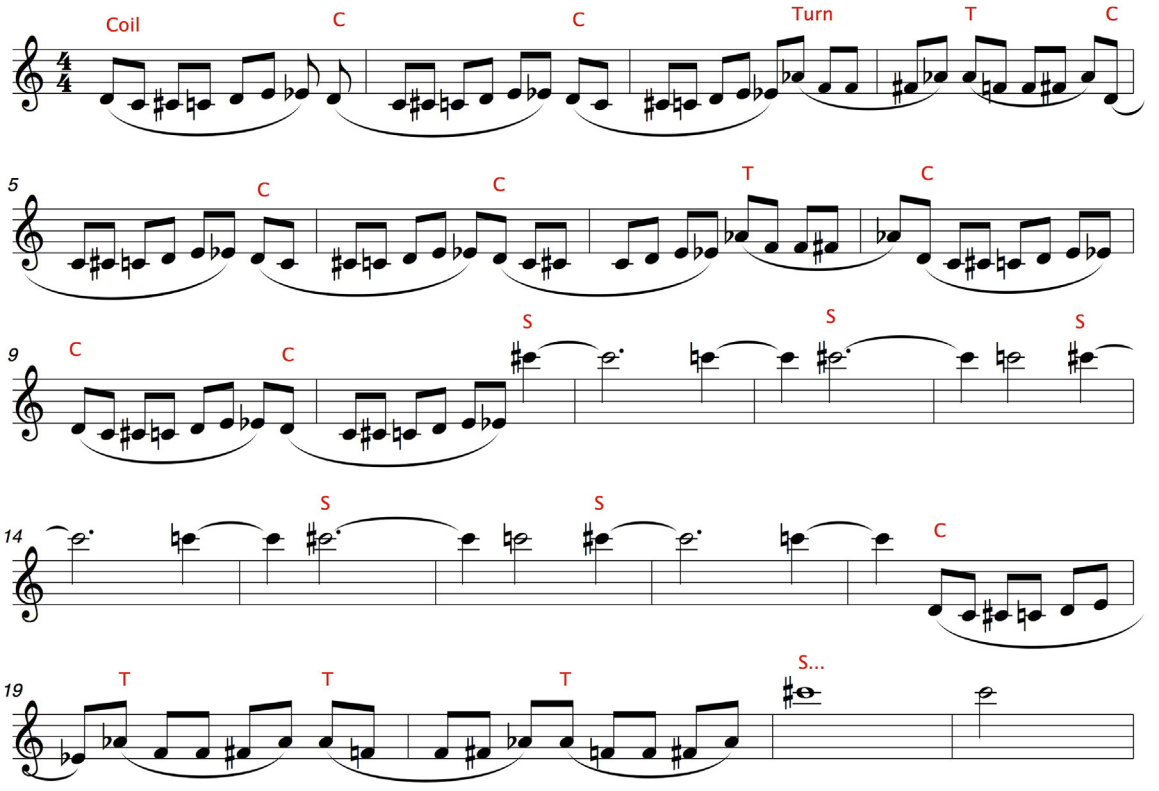

Fig. 3. Introductory excerpt from 1r75 score representing OHEL values.

characteristics of ascending and descending pitches (see Fig. 4 and supplementary Audio 4).

The OACC values were mapped by means of the natural log algorithm. The log algorithm was selected to add pitch space to the low and mid range values, giving them more separation for clarity and definition than the alternative proportionate mapping system with equal distribution. The log mapping system is ideal as it stretches the available low pitch space and narrows the available span for high pitch space to accommodate a large portion of low range data points for the protein

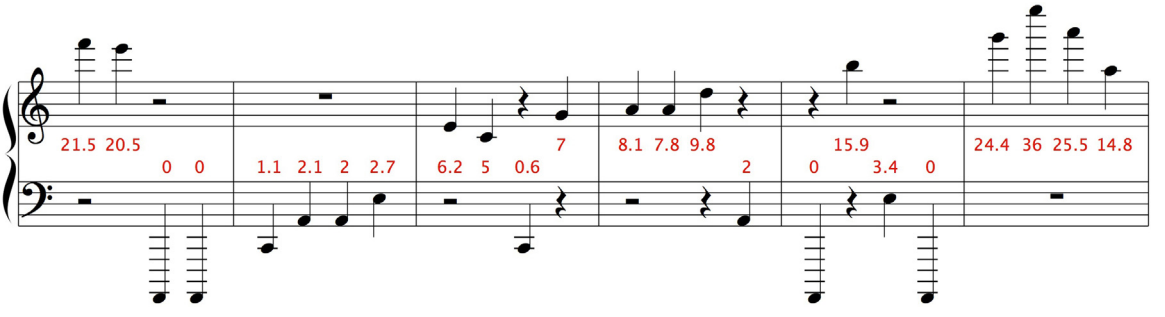

Fig. 4. Introductory excerpt from 4ake score representing OACC values.

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

structures selected. As an example, if we observe the 4ake OACC values we will see the data points reside in a numeric span of 0.00 to 54.27; however, only about 21 of 214 data points are above the value 30.0, so the preponderance of values are located between 0–30. This sonic profile with log mapping will enhance the individual characteristics of low OACC values in a manner that a Mercator projection will enhance the size of landmasses closer to the poles of the earth. Once the log mapping results were obtained, these pitches are slightly modified to adhere to a C major diatonic pitch collection. The added diatonicism creates a more listener-friendly environment than most chromatic representations.

3. Results

A perceptual survey with 38 human subjects was conducted to determine how well the protein sonifications are perceived in relation to their data sources. In the survey, participants were asked to compare sonifications with respective visual components (protein data or unrelated image). The human subjects responded by choosing one of four multiple choice responses for each image and accompanying sounds: A) similar, B) not similar, C) sometimes similar, D) mostly similar. This survey was granted exemption from federal regulations for the protection of human subjects (Title 45, Part 46.101(b)(1–6) by the Institutional Review Board for Human Subjects Research at Eastern Washington University (Review HS-4405). All participants were informed that the surveys were conducted on a voluntary basis, and the participants could leave at any time. No formal consent to participate in the survey was necessary.

3.1. Perceptual experiments

The survey comprised of six experiments seeking to establish the degree that people can correlate a visual image or graph of protein secondary structures with corresponding sonifications based on data-to-melodies. The testing began with two experiments dedicated to establishing controls for positive (similar) and negative (not similar) responses. The subjects were then tested with four experiments using our protein sonifications with traditional visual displays for AA, OACC, and OHEL. Two experiments were dedicated to the 1ny9 AA whole residue hydrophobicity sonification without scroll bar visual cues. Participants saw a standard graph representing the AA sequential order on the x-axis and whole residue hydrophobicity values on the y-axis; however, neither x nor y values or axes were labelled on the graph for the first AA experiment; subjects simply saw a graph line with dots and listened to a melody. In the second AA experiment, placed at the end of the testing period, subjects saw a line graph with dots and corresponding AA letters for each data point. Another experiment used a similar visual representation with a graph line only to represent 4ake OACC data. The sequential order of OACC data were featured on the x-axis and depth in a range of

 http://dx.doi.org/10.1016/j.heliyon.2016.e00175

0 to 21.54 was featured on the y-axis. Participants saw a simple graph line with dots. There were no labels and no scroll bar. In a final experiment we used a sequential list of OHEL data featured as one of the available RasMol 3D models. Participants saw a series of colored balls on the displayed model of the1r75 protein structure light up one at a time in sequence and in coordination with our OHEL sonification. Each color on the Rasmol 3D ball and stick model referenced an OHEL shape and each shape was referenced by corresponding musical motive. The synchronization of sound and light served as an alternative visual cue to the scroll bar in the controls.

3.2. Participants

- Most participants were students.
- Participants were generally between the ages of 18 and 60.
- Participants were asked if they had musical training and/or studied proteins.
- 21 participants (or 55%) have had more than a year of musical training.
- 17 participants (or 45%) have had up to a year or more of studies with proteins.
- 10 participants (or 26%) had neither musical training nor studies with proteins.

3.3. Conditions

- Sonification survey lasted no more than 20 min. The minimum testing time was 7 and half minutes.
- Participants were tested alone or in small groups under close observation.
- Visuals were observed from a large computer screen and sounds played from loud speakers.
- Both visuals and sound came from the same mp4 file source.
- Participants were not allowed to hear an example twice.
- Respondents were asked to choose the answer that best represents their perception when comparing the sounds with the image. Their choices were recorded on paper by checking empty boxes next to their selections.
- The order of multiple choice answers (as seen above) was uniform throughout the testing sessions.
- Questions for most of the experiments were presented with excerpts of our sonifications to reduce listener fatigue.

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

3.4. Controls

In a negative control, participants were given an image of four circles accompanied by a protein sonification that has no intended reference or connection to the circles (see supplementary Video 1). A synchronized scroll bar served as a visual guide from left to right for the duration of the sonification. 68% of participants perceived the image and sounds to be “not similar”. An additional 24% responded with “sometimes similar”. There was a 92% overall negative response rate. In one of two controls seeking a positive result, participants were given an image with a series of triangles increasing in height and accompanied by a sonification of numbers from Pascal's triangle (see supplementary Video 2). A synchronized scroll bar served as a visual guide from left to right for the duration of the sonification. Participants found a direct reference to the image, as 92% responded positively with “similar” (60.5%) or “mostly similar” (31.5%).

3.5. Survey results with supporting material and authors' comments

AA (1ny9) sample sonification with a line graph that plots high and low hydrophobicity points in the 1ny9 sequence, with no direct AA data references and no synchronized visual aid.

- 60% responded “similar”
- 15% responded “mostly similar”

Structure: 1ny9 ribbon structure (see Fig. 5) colored according to secondary residue type (green: hydrophobic (aliphatic), orange: hydrophobic (aromatic), blue: positively charged, red: negatively charged, magenta: polar but uncharged. Note how “green” and “orange” tend to point “inwards”)

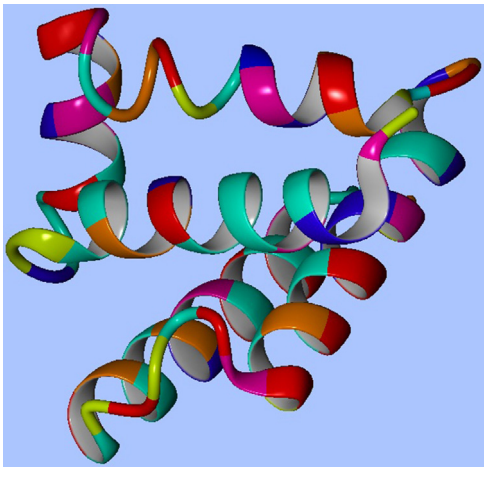

Fig. 5. 1ny9 ribbon structure.

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

Musical Score: See Supplementary Figure 1 and Supplementary Audio 2.

Input data: Supplementary Table 1.

Authors' comments: This is received as a melody that is easily reconcilable with the AA rendering of the structure (hydropathy) but the listener needs to be able to visualize the structure while listening. Charged amino acids (K, D, E and R) and aromatic amino acids (W and F) are most discernible due to their position on the extreme ends of the octanol scale (after averaging) used here. In this model, the extremities of a given tessitura with high or low pitches, provide salient perceptual separation. The majority of amino acids, however, are found in the mid range of the octanol scale [19] and these are notably hard to differentiate in relatively close proximity when clustered in a two-octave sonification span. Hence, in relation to pitch mapping, the mid range amino acids are far less salient than those amino acids on the extreme ends. The mid range amino acids might be relatively discernible according to some conditions such as, the listener has perfect pitch, the tempo is very slow, the mid range amino acids in the sequence are reduced in number, or the mid range amino acids are repeated. Further research would be required to improve the ability to discern mid range AA values.

AA (1ny9) sample sonification (same as above) with a line graph plotting high and low hydrophobicity points with more vertical space and protein sequence labels: W, Q, R, I...., but no synchronized visual aid.

- 55% responded “similar”
- 25% responded “mostly similar”

Authors' comments: Eighty percent of participants demonstrated the ability to correlate pitch height and melodic direction with equivalent contours of graph lines via auditory perception. In this example, pitch is a significant sonification element.

OHEL (1r75) sample sonification with colored balls that light up with each sonified musical motive as a visual aid.

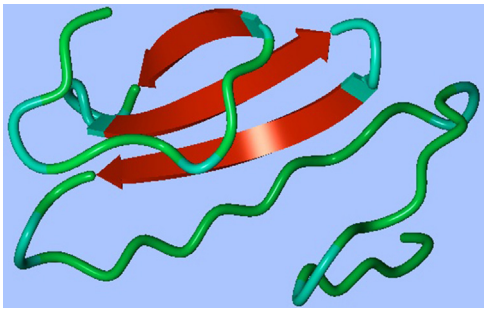

Fig. 6. 1r75 ribbon structure.

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

- 42% responded "similar"
- 44% responded "mostly similar"

Structure: 1r75 ribbon structure (see Fig. 6) colored according to secondary structure (red: β-strand, cyan: "unstructured" regions).

Musical score: See Supplementary Figure 2 and Supplementary Audio 3.

Input data: Supplementary Table 2.

Authors' comments: In this example the OHEL shapes (H, C, T, S) are mapped to create musical motives (or melodic fragments) of corresponding shapes. The melodic result is essentially a string of repeating musical motives with some rhythmic variety. The composite, or overall melodic result reflects the secondary structure in a way that makes stretches of different secondary structure easily recognizable – in particular, the eight stretches of β-strand that comprise the main corpus of this protein. Proteins in different fold classes, for example all-α proteins or α/β proteins, sound very different in a characteristic and memorable way. The music is somewhat laboured, which paradoxically enhances recognition but is not necessarily conducive to making enjoyable listening. Conclusions drawn from the survey are as follows: The sonified data shapes are easy to track with some training, since there are only a few shapes to listen for, and each shape is unique. In the survey, however, colored balls were used for data representation, where each OHEL shape had a corresponding color. This level of abstraction and the circuitous trajectory of the protein led us to rely on a synchronized visual aid to guide the listeners. The 86% positive response rate (from "similar" and "mostly similar" responses) reflects the success of the sonification and the visual aid combined, however the "similar" response rate, when measured alone, was one of the lowest in the survey. It is possible that the connection between varying musical motives and the uniform shape of the balls weakened the perceived similarities, hence it might be more advantageous to visualize the actual OHEL shapes with the aid of a suitable computer graphics or virtual reality device to match while the motives are played. The lower than expected "similar" response rate may also be due to the consequences of stringing together motives of varying lengths, which in some instances contributes to a perceived metric displacement and metric ambiguity. The absence of a consistent metric pattern, in the context of uniformly spaced balls, may have confused some listeners, especially among those with musical backgrounds.

OACC (4ake) sample sonification with a line graph containing no specific data references and no synchronized visual aid

- 47% responded "similar"
- 26% responded "mostly similar"

 http://dx.doi.org/10.1016/j.heliyon.2016.e00175

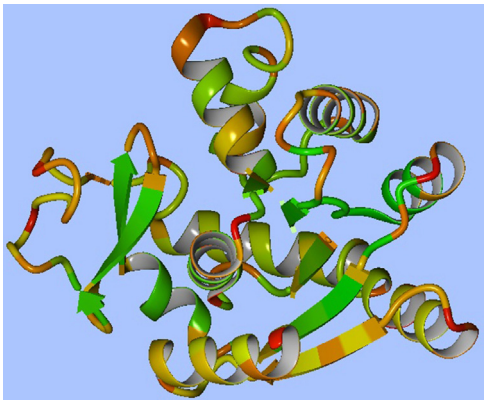

Fig. 7. 4ake ribbon structure.

Structure: 4ake ribbon structure (see Fig. 7) colored according to accessibility (green: buried, orange: surface).

Musical Score: See Supplementary Figure 3 and Supplementary Audio 4.

Input data: Supplementary Table 3.

Authors' comments: Seventy-three percent of participants demonstrated the ability to correlate pitch height and melodic direction with equivalent contours of graph lines via auditory perception. The way this sonification was designed led to a melody with extremely high and low pitches that clearly gave a sense of "commuting" between the interior of the protein (green color in Fig. 7) and its surface (orange in Fig. 7). While the sonification gives a sense of the "diving" and "resurfacing" that a polypeptide chain indeed does, it mainly gives clues as to how the protein is constructed and it is unlikely that this approach will make it easy to compare different proteins.

3.6. Summary

The sonification response results from the survey demonstrate that a significant majority of listeners can discern correlations between data related images and sounds when pitch is a primary mapping parameter (see Table 2). Furthermore, visual aids such as synchronized scroll bars are helpful, but not essential. This suggests that pattern recognition of protein structures can be attained on an intuitive level with little training.

Availability of data to music software

Instructions on how to make one's own melodies from data can be obtained from the "How To" page at: http://www.musicalgorithms.org (see version 3.2)

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

Table 2. Perceptual sonification survey results. OHEL and controls used real time visual cues, while AA and OACC sonification excerpts used no real time visual cues. Participants were able to perceive similarities between visual representations and audible melodies generated by the same data.

Participants Sonification Perceptual Survey (Human Subjects 1-38)	Control Anticipating Positive ("A" or "D" Responses)	Control Anticipating Negative ("B" or "C" Responses)	AA 1ny9 excerpt	OACC 4ake excerpt	OHEL 1r75 complete	AA 1ny9 in detail
Subject #1	C	B	A	A	D	N/A
Subject #2	A	B	A	A	A	N/A
Subject #3	A	B	C	B	D	C
Subject #4	A	B	C	C	A	C
Subject #5	D	C	D	A	A	A
Subject #6	D	B	A	D	D	D
Subject #7	D	B	D	D	D	A
Subject #8	D	A	C	A	D	D
Subject #9	A	B	D	A	D	D
Subject #10	A	D	A	A	C	C
Subject #11	D	B	A	A	D	A
Subject #12	A	B	A	D	A	D
Subject #13	A	B	D	D	A	A
Subject #14	A	C	A	D	A	A
Subject #15	A	C	A	A	D	A
Subject #16	A	C	D	D	A	A
Subject #17	D	A	B	D	D	D
Subject #18	A	C	A	D	D	C
Subject #19	D	B	B	C	A	D
Subject #20	A	B	A	B	A	D
Subject #21	A	B	A	A	A	A
Subject #22	D	B	B	B	C	C
Subject #23	D	B	A	D	D	A

(*Continued*)

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

Table 2. (*Continued*)

Participants Sonification Perceptual Survey (Human Subjects 1-38)	Control Anticipating Positive ("A" or "D" Responses)	Control Anticipating Negative ("B" or "C" Responses)	AA 1ny9 excerpt	OACC 4ake excerpt	OHEL 1r75 complete	AA 1ny9 in detail
Subject #24	A	B	C	C	C	A
Subject #25	A	B	C	C	D	A
Subject #26	A	B	A	A	D	A
Subject #27	A	B	A	A	A	A
Subject #28	D	B	A	A	A	A
Subject #29	C	B	A	A	A	A
Subject #30	A	C	A	A	D	A
Subject #31	A	C	A	D	C	D
Subject #32	A	B	D	C	A	C
Subject #33	A	B	A	A	A	A
Subject #34	A	C	A	A	A	A
Subject #35	D	C	A	A	D	A
Subject #36	D	B	A	C	C	D
Subject #37	C	B	B	B	D	B
Subject #38	A	B	A	A	D	A
Total Percentages of Response Types						
A Similar	60.53%	5.26%	60.53%	47.37%	42.11%	55.56%
B Not Similar	0.00%	68.42%	10.53%	10.53%	0.00%	2.78%
C Somewhat Similar	7.89%	23.68%	13.16%	15.79%	13.16%	16.67%
D Mostly Similar	31.58%	2.63%	15.79%	26.32%	44.74%	25.00%

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

4. Discussion

We have suggested several ways in which protein sequences can be "heard" rather than "read". The amino acid sequence is rendered as a string of musical notes with sound qualities that reflect the properties of these residues. There are several different ways of encoding these properties and we have examined several of these, all with interesting and memorable, but different outcomes. Certain of these (OACC/PACC and AA) possess the property of having a 1:1 ratio between pitches and residues, while OHEL/PHEL has been constructed specifically in order to capture the defining structural feature of each secondary structure type (e.g. 2 pitches for S/E, 3 pitches for 3_{10}, and 4 pitches for α-helix).

By listening to a melodic shape for a given protein sequence, an impression of its structure can be built up. We propose this as an excellent learning tool for those wishing better to understand and discriminate protein fold structures, and as an aid to fold identification itself. We intend to extend this work towards a more complete domain fold identification system, so that with practice, and by comparison with the sounds corresponding to other protein types, this can lead to a way of recognizing the 3D folds of different proteins. Further, we are aware that proteins are flexible entities that typically switch between two different structures in the course of exercising their function [22]. It would be interesting if our musical approach could help to identify the signatures corresponding to the residue positions that are responsible for this particularly important protein function. The musical patterns are complex, and this will not only be because protein folds are complex but also due to the need to switch between different structural states. The primary sequence of proteins has to cater for more than folding and switching between folds, but for a whole range of other functions [22] including how the protein arrives at its destination in (or outside) the cell, which requires the attachment of post-translational signals on its surface. These functions are also encoded genetically [22]. For these reasons we expect there to be many developments in the area where proteomics and sonification overlap.

We should mention that similar attention is being paid to sonification of DNA sequences and gene expression [23, 24]. These also encapsulate structural information but in a somewhat different way, the protein sequences are more closely related to the events that actually take place at the phenotypic level in the living cell. These protein sequences are encoded in DNA but the latter can also encompass regions that have to do with cell differentiation and epigenetic control [24] which are outside of and distinct from the protein coding regions. Thus there is every reason to engage in sonification studies at the DNA level also.

http://dx.doi.org/10.1016/j.heliyon.2016.e00175

Declarations

Author contribution statement

Robert P. Bywater, Jonathan N. Middleton: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This work was supported by an Eastern Washington University Faculty Research and Creative Works Summer Grant. This work was also supported in part by Tekes – The Finnish funding agency for innovation (decision 40296/14).

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at 10.1016/j.heliyon.2016.e00175.

Acknowledgements

The authors would like to acknowledge the support of Eastern Washington University and the University of Tampere for helping us move sonification research forward in the context of biological sciences.

References

[1] G. Bellesia, A.I. Jewett, J.E. Shea, Sequence periodicity and secondary structure propensity in model proteins, Protein Sci. 19 (2010) 141–154.

[2] S. Hovmöller, T.P. Zhou, T. Ohlson, Conformations of amino acids in proteins, Acta Cryst. D 58 (2002) 768–776.

[3] R.P. Bywater, D. Thomas, G. Vriend, A sequence and structural study of transmembrane helices, J. Comput-Aided Mol. Des. 15 (2001) 533–552.

[4] R.P. Bywater, V. Veryazov, The preferred conformation of dipeptides in the context of biosynthesis, Naturwissenschaften 100 (2013) 853–859.

[5] A.G. Murzin, S.E. Brenner, T. Hubbard, et al., SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (1995) 536–540.

[6] C.A. Orengo, A.D. Michie, S. Jones, et al., CATH – a hierarchic classification of protein domain structures, Structure 15 (1997) 1093–1108.

[7] N. Koga, R. Tatsumi-Koga, G.H. Liu, T.B. Acton, G.T. Montelione, D. Baker, et al., Principles for designing ideal protein structures, Nature 491 (2012) 222–227.

[8] S. Barrass, P. Vickers, Sonification design and aesthetics, The Sonification Handbook, Logos Verlag, 2011, pp. 145–172.

[9] B.N. Walker, G. Kramer, Mappings and metaphors in auditory displays: an experiment assessment, ACM Trans. Appl. Percept. 2 (4) (2005) 407–412.

[10] The Sonification Handbook, In: T. Hermann, A. Hunt, J.G. Neuhoff (Eds.), Logos Verlag, 2011.

[11] F. Dombois, G. Eckel, Audification, The Sonification Handbook, Logos Verlag, 2011.

[12] J. Dunn, M.A. Clark, Life music: The sonification of proteins, Leonardo 32 (1) (1999) 25–32.

[13] R. Takahashi, J.H. Miller, Conversion of amino-acid sequences in proteins to classical music: Search for auditory patterns, Genome Biol. 8 (2007) 405.

[14] M.A. Garcia-Ruiz, J.R. Guiterrez-Pulido, An overview of auditory display to assist comprehension of molecular information, Interact. Comput. 18 (2006) 853–868.

[15] A. Supper, Sublime frequencies: the construction of sublime listening experiences in the sonification of scientific data, Soc. Stud. Sci. 44 (1) (2014) 34–58.

[16] P. Minary, M. Levitt, Probing protein fold space with a simplified model, J. Mol. Biol. 375 (2008) 920–933.

[17] G. Vriend, WHAT IF: a molecular modelling and drug design program, J. Mol. Graph. 8 (1990) 52–56.

[18] B. Rost, G. Yachdav, J.F. Liu, The PredictProtein server, Nucl. Acids Res. 32 (2004) W321–W326.

[19] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637.

[20] W.C. Wimley, S.H. White, Membrane partitioning: Distinguishing bilayer effects from the hydrophobic effect, Biochemistry 32 (1993) 6307–6312.

 http://dx.doi.org/10.1016/j.heliyon.2016.e00175

[21] J. Middleton, D. Dowd, Web-based algorithmic composition from extramusical resources, Leonardo 41 (2) (2008) 128–135.

[22] R.P. Bywater, Protein folding: a problem with multiple solutions, J. Biomol. Struct. Dyn. 31 (4) (2013) 351–362.

[23] M.S. Staege, A short treatise concerning a musical approach for the interpretation of gene expression data, Sci. Rep. 5 (2015) 15281.

[24] D. Brocks, Musical patterns for comparative epigenomics, Clin. Epigenetics 7 (2015) 94.

 http://dx.doi.org/10.1016/j.heliyon.2016.e00175